Available Online: 26 June 2026 DOI: 10.54254/2977-3903/2026.34776

The progress and application of multimodal deep learning artificial intelligence technology in personalized emotional companionship

Jingru Geng

College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China

2025014212@zwu.edu.cn

Abstract. At present, artificial intelligence technology can achieve simple communication with users and provide certain answers to the questions raised by users. With the vigorous development of artificial intelligence technology, obtaining emotional companionship through artificial intelligence technology has become a trend and trend. This study focuses on the application of multimodal artificial intelligence technology in emotional companionship, especially in providing personalized emotional companionship. It summarizes relevant literature and research results in recent years, and summarizes the current situation and problems of artificial intelligence technology in achieving personalized emotional companionship under multimodal conditions. It provides reference for solving problems such as difficulty in analyzing and extracting information from historical information. Research has found that at present, artificial intelligence technology can better understand user expressions and provide positive feedback in personalized emotional companionship, but there are still drawbacks such as "forgetting memories", "awkward conversations", and insufficient interaction effects.

Keywords: artificial intelligence, emotional companionship, personalization, deep learning, multimodal

1. Introduction

Emotional companionship, as an effective way to alleviate emotional problems, has a wide range of applications in daily life. Compared to communicating with professional psychological counselors, people tend to seek emotional companionship in artificial intelligence technology, and the rapid development of artificial intelligence technology provides strong technical support for this. There are many related achievements at present, such as chat gpt, Replika, and the application of big models such as bean buns, which can provide certain answers to users' questions and help alleviate negative emotions. However, there are currently problems with "machine companions" such as difficulty in remembering historical conversations. This article summarizes the achievements of multimodal development and related issues, providing a research foundation for achieving strong artificial intelligence to provide personalized emotional companionship for humans. Firstly, the article introduces the basic artificial intelligence technology. The architecture principles and composition mechanisms of neural networks. Here, techniques such as convolutional neural networks, recurrent neural networks, long short-term memory networks, and residual neural networks are successively explained. This is the foundation for achieving personalized emotional companionship. Subsequently, some meaningful models were proposed that can solve the problems encountered in development. In the text recognition module, the original technology utilizes the bag of words model, which is prone to overfitting problems in machines under this model. In order to improve the accuracy of recognition, pre trained GloVe layers and Bidirectional Long Short-Term Memory layers (BiLSTM) are used [1]. Subsequently, the analysis of technical features at different levels helps to gain a detailed understanding of user intent and depict user profiles. When the model is trained, it is trained using a large amount of data from other users, and then optimized by comparing it with the target user; Another method is to mix other user data with the target user data for training. The above two methods effectively enhance the model's portrayal of user image, reduce training frequency and cost, and achieve effective personalized emotional companionship [2]. In terms of interaction methods, different interaction methods need to be set for different user groups and needs. Finally, this article analyzes the shortcomings and challenges of existing technologies and provides certain research directions. In terms of language and text recognition, due to the inaccuracy of language and text, the robustness of the recognition text model still needs to be improved. In terms of data collection, different countries and regions have different usage norms for data, which also limits the formation of unified evaluation standards. The accuracy of artificial intelligence technology's answers still needs to be improved, and it may even lead users to exhibit extreme behavior, which forces relevant departments to enact corresponding laws and regulations to face the current risks.

2. Characteristics of emotional companionship in deep learning and existing technologies

Deep learning technology is machine learning based on deep neural networks that can extract information from data. Multimodal refers to the use of multiple information carriers in the input and output stages. Emotional companionship is the understanding and guidance of one party's immediate emotional state towards another. Unlike other forms of assistance, emotional companionship focuses more on understanding and supporting rather than judging right from wrong and solving problems. This means that personalized services that better meet the needs of users are needed in emotional companionship. Current technologies such as voice interaction, image recognition, and large language models can meet some of the needs of users.

Convolutional Neural Network (CNN) is a simulation of the human brain's processing of images, mainly composed of input layer, convolutional layer, pooling layer, and fully connected layer. One of the characteristics of convolutional neural networks is their strong robustness in image processing such as translation, rotation, and scaling [3]. Recurrent Neural Networks (RNNs) add connections W_s between layers on the basis of convolutional neural networks, which not only reduces training time but also enables machine learning to understand context and achieve better language comprehension. However, the implementation of memory function in recurrent neural networks requires continuous data input, and the data is constantly updated, which can easily lead to forgetting.

The hidden state (H_t) in recurrent neural networks is determined by the current input x_t and the previous hidden state h_{t-1} . The hidden state can be used as input for the next step, making it have memory function and helping to understand contextual content.

$$h_t = \sigma(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{1}$$

Traditional RNN's hidden state transfer:

$$h_t = tanh(W \cdot h_{t-1} + U \cdot x_t) \tag{2}$$

During gradient backpropagation:

$$\prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}} = \prod_{j=k+1}^{t} tanh\prime(W_s) \cdot W_s \tag{3}$$

Among them, $tanh'(x) \in (0, 1]$ and usually less than 1; If the spectral norm of the weight W_s is less than 1, the multiplication result will exponentially approach 0, leading to the disappearance of the gradient. The Long Short Term Memory network (LSTM) incorporates more matrices into the model through gating mechanisms and cell state design, solving the gradient vanishing problem in traditional RNN techniques. The forget gate in the gating mechanism determines how much history is retained, while the input gate determines how much new information is retained, as shown in Figure 1.

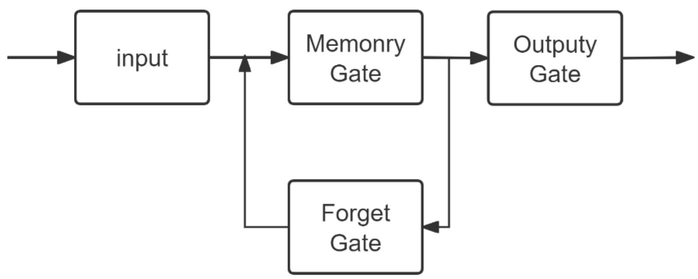

Figure 1. Architecture of gate control mechanism for long short term memory network

The core of LSTM is the cell state C_t , which can linearly transmit information between each time step, avoiding the problem of gradient exponential decay in RNN.

The comprehension ability of neural networks does not increase with the depth of the network. Residual Neural Network (ResNet) adds residuals, which are the new parts that need to be learned, on the basis of the original learning, so that the original ability can be retained and further deepened. Residual neural networks increase the depth of neural networks to prevent collapse, which is an improvement over previous neural networks.

$$F_{l+1} = \sigma(W_l F_L + b_l) + F_l \tag{4}$$

Generative Adversarial Networks (GANs) consist of generators and discriminators. The generator model generates a large number of data samples, and then mixes these generated data with accurate input data. The discriminator model determines the authenticity of these data. Two models continuously learn through zero sum games. The data generated by generative adversarial networks can be used for model training.

All images in the main text must ensure sufficient clarity to clearly display the content within the images. All formulas, tables, and program codes in the article must be in an editable format and should not be presented as photos or screenshots. All references must be cited in the main text.

3. Technology update

3.1. Model layer

In terms of text recognition, traditional machine learning often builds bag of words models to measure the importance of terms in corpora. For this purpose, the basic model of neural network architecture consists of an embedding layer, a recurrent layer, and a dense layer. In this architecture, due to the inclusion of a large number of parameters, it is not suitable for computing small datasets in emotions, which can easily lead to severe overfitting. To optimize this problem, a pre trained GloVe layer is used. To improve the accuracy of recognition, a Bidirectional Recurrent Layer (BiLSTM) is used to capture more semantic information [1].

A fine-tuning model is proposed to avoid overfitting during model training. Firstly, pre training with a large amount of other user data, and then retraining with target user data, can avoid overfitting and better form personalized models [2].

Similar to fine-tuning models, hybrid models differ in that they mix a large amount of other user data with the target user data for unified training, which only trains once and reduces training costs [2]. The fusion of multi head self attention mechanism and position encoding technology with eye tracking modality significantly improves the accuracy of distinguishing between users at risk of depression and those in a healthy state.

In terms of speech recognition, a MCDNet multi-channel one-dimensional convolutional neural network has been proposed to improve recognition accuracy and robustness. This model consists of a Feature Extraction Module (FEB), a Feature Learning Module (FLB), and a Classification Module (CB). This model has improved the accuracy of emotion recognition through speech compared to the original model [4].

3.2. Data layer

The data is preprocessed during collection (including tail reduction, filtering, downsampling, normalization, and segmentation), and personalized classification processing is applied to different datasets of each user, significantly improving recognition performance [2]. The raw signal contains significant spike noise, baseline drift, and high-frequency interference. By using Winsorization, extreme values outside the range are reduced to a specified range, preserving data features while removing spikes. In order to remove high-frequency noise and obtain a smooth signal, the data is processed again using a filtering process. Downsampling can reduce data volume and save storage space while preserving key information. The function of normalization is to scale the signal to the interval of [0,1], which is beneficial for subsequent modeling. Then divide the signal into different segments and label them to complete segmentation&labeling. Figure 2 shows the process of data preprocessing and the characteristics of each step.

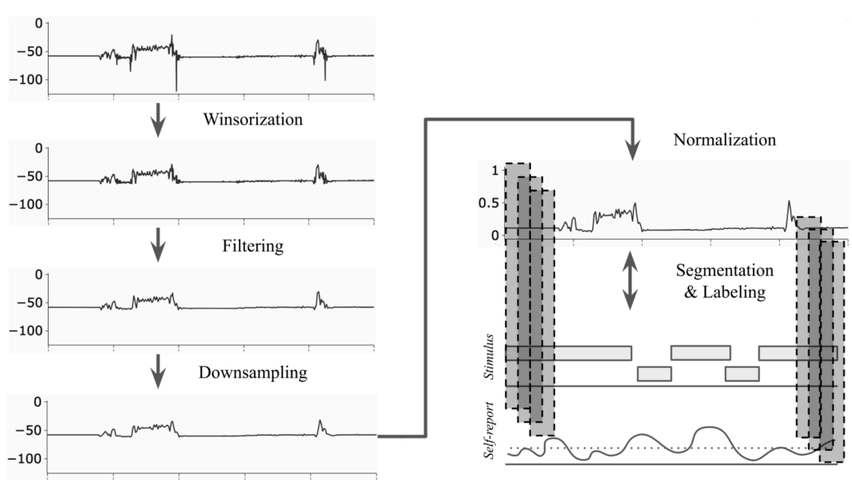

Figure 2. Dataset preprocessing operation architecture diagram [3]

After preprocessing, the extreme special values of the signal are removed, while the signal features are preserved. Preprocessing not only saves storage space, but also improves the accuracy of machine recognition of speech signals.

3.3. Interaction layer

Generative interaction is superior to retrieval based interaction; Multimodal interaction is superior to text-based interaction; The interaction mode on the mobile end is superior to that on the web version [5]. Reciprocal self disclosure is a communication method that automatically shares one's own experiences when the other person is under pressure during a conversation. This approach can effectively alleviate people's anxiety [6]. Dialogue based artificial intelligence can provide users with a non judgmental empathetic support [7]. Providing 24-hour uninterrupted service to meet users' emotional companionship needs is also a performance of artificial intelligence that surpasses human experts [8]. When users seek emotional support and assistance, artificial intelligence that enhances dialogue humor can often provide better emotional companionship [8].

3.4. Domain datasets and evaluation standards

The AI Dialogue Robot (AI-CA) model can effectively alleviate depressive symptoms and psychological distress, and the generative multimodal mobile design approach is more effective for clinical and elderly populations [5]. The frequency of users using artificial intelligence is closely related to the emotions they perceive, and users with higher usage frequency have significant improvements in interpersonal communication, emotional support, and other aspects. In terms of user trust, artificial intelligence that creates a trust environment can gain the favor of users [8]. In terms of constructing domain datasets, a standardized data resource system covering multiple modalities and diseases has been established. The training included data from 175 different groups of subjects, including a healthy control group and a depression risk group, to ensure that the model had not been exposed to relevant data from the test subjects before training [3]. Judging the psychological state of users is an indispensable task in the evaluation standard system. For the classification task of whether users have psychological disorders, the multi-modal fusion self attention network achieved a classification accuracy of 93.62%, significantly surpassing the traditional baseline model [3].

4. Challenges and prospects

4.1. The problems that still exist at this stage

Due to the inaccuracy of language and writing, as well as factors such as user language habits and cultural backgrounds, the robustness of text recognition models still needs to be improved [1]. When users use expressions such as dialects, metaphors, or slang, or describe emotional experiences in different cultural backgrounds, existing natural language processing models often struggle to accurately capture the user's true psychological state, leading to bias in recognizing user intentions or distortion in emotional analysis, which may directly affect the accuracy and effectiveness of intervention recommendations. Privacy protection and other security issues affect the assessment and standards of mental health. In the process of collecting and processing sensitive psychological data of users, how to ensure data transmission encryption, storage desensitization, and access permission hierarchical management has become a key bottleneck restricting the construction of large-scale datasets and model training. In addition, there are significant differences in the usage norms, informed consent requirements, and cross-border data flow restrictions for users' mental health data in different countries and regions, which poses complex challenges in establishing unified and comparable evaluation standards. The lack of standardized privacy protection mechanisms not only limits the acquisition of high-quality annotated data, but also makes it difficult to achieve consistency in the deployment and evaluation of models in different jurisdictions. Artificial intelligence provides confusing answers and sometimes even guides users to take extreme actions, causing security risks [9]. When the model faces emergency situations such as high suicide risk, self harm ideation, or severe mental crisis, if it lacks effective safety alignment mechanisms and risk identification capabilities, it may generate logical contradictions, emotional indifference, or potentially harmful responses. This "illusion" phenomenon is particularly dangerous in open dialogue, as users are in a period of psychological vulnerability and have a high level of trust in AI advice. Improper guidance may exacerbate their sense of despair or impulsive behavior. Therefore, establishing special safety barriers, crisis identification protocols, and manual takeover mechanisms for the field of mental health has become an indispensable component of current evaluation standards and urgently needs to be included in the core indicator system of model robustness assessment.

4.2. Outlook suggestions

In terms of model training, a large number of professional talents are needed, and there is an urgent need to face the talent gap in the corresponding field. The quantity and quality of data greatly affect the training effectiveness of the model, so techniques such as generative adversarial networks and transfer learning can be used to expand the data sample [3]. Adding images with lower accuracy for model training can improve the model's recognition ability for sketches and other images [10]. Generative AI technologies have effectively enhanced the intervention capabilities for mental health. Implementing personalization requires a certain amount of unknown user data, and user participation in model training is extremely important [2]. At the user interaction level, intelligent agents need to master some conversational techniques when outputting, such as incorporating reciprocal self disclosure when providing emotional support, which is more conducive to users alleviating negative emotions. In terms of security, artificial intelligence should be subject to supervisory measures to ensure user privacy and security. At the design level, increasing user engagement can effectively enhance personalization and meet the different needs of different groups of people [8]. In terms of cross-cultural adaptability, existing datasets have covered samples from multiple regions such as China, the United States, Europe, and Australia, but cultural diversity still needs to be improved. Future dataset construction needs to further focus on special scenarios such as low - and middle-income countries and conflict areas to enhance the global applicability of the model.

5. Conclusion

This article systematically reviews the dataset construction and evaluation system of multimodal deep learning in personalized emotional companionship, focusing on multimodal data such as text and speech. Briefly introduced the existing neural network technology. To solve the problems of insufficient memory, strong fitting, and inaccurate recognition in practical use, relevant innovative models and optimization solutions have been proposed successively, effectively improving the accuracy and adaptability of emotion recognition, and promoting the development of personalized emotional companionship towards a more accurate and stable direction. Finally, the existing problems were analyzed, including deficiencies in model training, user interaction, security, and cross-cultural adaptability. Generative, mobile, and multimodal fusion schemes have better effects. The problems of insufficient model robustness, privacy security, and ethical risks still exist.

Future researchers need to continuously improve relevant technologies and regulations in terms of data quantity and quality, model optimization, interaction mode selection, and normative value orientation.

References

[1] Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems, 115*, 24–35. https: //doi.org/10.1016/j.dss.2018.09.002

[2] Han, Y., Zhang, P., Park, M., & Lee, U. (2024). Systematic evaluation of personalized deep learning models for affect recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8*(4), Article 206, 1–35. https: //doi.org/10.1145/3699724

[3] Zhu, F. (2025). *Research on multi-mode fusion emotional disorder detection algorithm based on deep learning* [Doctoral dissertation, University of Chinese Academy of Sciences]. https: //doi.org/10.27605/d.cnki.gkxgs.2025.000009

[4] Liu, M. (2024). *Research on speech emotion recognition method based on convolutional neural network* [Master's thesis, Shantou University].

[5] Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine, 6*(1), Article 236.

[6] Meng, J., & Dai, Y. (2021). Emotional support from AI chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication, 26*, 207–222.

[7] Chandra, M., Hernandez, J., Ramos, G., Ershadi, M., Bhattacharjee, A., Amores, J., Okoli, E., Paradiso, A., Warreth, S., & Suh, J. (2025). *Longitudinal study on social and emotional use of AI conversational agent.* arXiv. https: //doi.org/10.48550/arXiv.2504.14112

[8] Chaudhry, B. M., & Debi, H. R. (2024). User perceptions and experiences of an AI-driven conversational agent for mental health support. *mHealth, 10*, Article 15.

[9] Adam, D. (2025). Supportive? Addictive? Abusive? How AI companions affect our mental health. *Nature, 641*(8062), 296–298.

[10] Ballester, P., & Araújo, R. M. (2016). On the performance of GoogLeNet and AlexNet applied to sketches. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI'16) (pp. 1124–1128). AAAI Press.